Title
Efficient Software for Multi-marker, Region-Based Analysis of GWAS Data.

Journal
G3 (Bethesda, Md.), 6(4)

ISSN
2160-1836

Authors
Sanjak, Jaleal S
Long, Anthony D
Thornton, Kevin R

Publication Date
2016-04-01

DOI
10.1534/g3.115.026013

Peer reviewed

Efficient Software for Multi-marker, Region-Based Analysis of GWAS Data

Jaleal S. Sanjak,*,†,1 Anthony D. Long,*,† and Kevin R. Thornton*,†

*Department of Ecology and Evolutionary Biology, and †Center for Complex Biological Systems, University of California Irvine, California 92697

ABSTRACT Genome-wide association studies (GWAS) have associated many single variants with complex disease, yet the better part of heritable complex disease risk remains unexplained. Analytical tools designed to work under specific population genetic models are needed. Rare variants are increasingly shown to be important in human complex disease, but most existing GWAS data do not cover rare variants. Explicit population genetic models predict that genes contributing to complex traits and experiencing recurrent, unconditionally deleterious, mutation will harbor multiple rare, causative mutations of subtle effect. It is difficult to identify genes harboring rare variants of large effect that contribute to complex disease risk via the single marker association tests typically used in GWAS. Gene/region-based association tests may have the power detect associations by combining information from multiple markers, but have yielded limited success in practice. This is partially because many methods have not been widely applied. Here, we empirically demonstrate the utility of a procedure based on the rank truncated product (RTP) method, filtered to reduce the effects of linkage disequilibrium. We apply the procedure to the Wellcome Trust Case Control Consortium (WTCCC) data set, and uncover previously unidentified associations, some of which have been replicated in much larger studies. We show that, in the absence of significant rare variant coverage, RTP based methods still have the power to detect associated genes. We recommend that RTP-based methods be applied to all existing GWAS data to maximize the usefulness of those data. For this, we provide efficient software implementing our procedure.

KEYWORDS
GWAS
gene-based rare variants

Revealing the genetic basis of common human diseases, such as diabetes and heart disease, remains a central challenge in human genetics. Family-based and twin-based studies estimate that the genetic component of disease risk is typically large. Genome-wide association studies (GWAS) have identified many genetic variants associated with complex human diseases (Welter *et al.* 2014), yet the heritability explained by specific statistically significant variants remains small in comparison to the total heritability estimates (Manolio *et al.* 2009; Visscher *et al.* 2012a). Various hypotheses explaining the "missing heritability problem" exist (Manolio *et al.* 2009; Visscher *et al.* 2012a; Gibson 2012; Robinson *et al.* 2014). Gene-by-gene, gene-by-environment, and other complex epistatic interactions might create statistical challenges for the detection of causal variants (Eichler *et al.* 2010; Wei *et al.* 2014), or might inflate total heritability estimates (Zuk *et al.* 2012). The missing heritability could be attributable to many common well-tagged variants that do not reach statistical significance because of their miniscule effect sizes (Fisher 1930; Visscher *et al.* 2008). Rare variants with large effects (RALE) might drive heritability and escape detection because they are not well-tagged by current genotyping methods (McClellan and King 2010; Cirulli and Goldstein 2010). Quantifying the roles of these nonmutually exclusive hypotheses is important for the design of future studies, and the development of new analytical tools (Visscher *et al.* 2012b). We still do not know exactly how mutational effect sizes underlying specific diseases map onto the human site-frequency spectrum. However, it is becoming increasingly clear that rare variants are an important contributor to the genetic basis of complex diseases (Auer *et al.* 2015; Prescott *et al.* 2015; Wessel and Goodarzi 2015; Purcell *et al.* 2014; Cruchaga *et al.* 2014; Huyghe *et al.* 2013; Nelson *et al.* 2012; Johansen *et al.* 2011).

doi: 10.1534/g3.115.026013
Manuscript received December 11, 2015; accepted for publication February 6, 2016; published Early Online February 9, 2016.

Supplemental Material is available online at www.g3journal.org/lookup/suppl/doi:10.1534/g3.115.026013/-/DC1
[1]Corresponding author: 321 Steinhaus Hall, University of California Irvine, CA 92697-1075. E-mail: jsanjak@uci.edu

The RALE hypothesis is particularly appealing to some because it is a prediction that arises naturally from population-genetic models of mutation-selection balance (Haldane 1927). Specifically, it arises from a model in which equilibrium allele frequencies and phenotypic effect sizes both reflect a balance between two things: recurrent unconditionally deleterious mutations occurring in a disease gene, and their elimination by natural selection (Pritchard 2001). A previous simulation study (Thornton *et al.* 2013) investigated a novel model where standing quantitative genetic variation in complex disease genes of large effect is maintained via partially noncomplementing mutations. An important prediction of this model is that a gene region can harbor several, individually rare, variants which all contribute to a complex disease phenotype. Such allelic heterogeneity is predicted to pose complications for genome wide association studies (McClellan and King 2010). In particular, we know that single-marker association tests do not have sufficient statistical power in these cases (Johnston *et al.* 2015; Sham and Purcell 2014; Spencer *et al.* 2009). Further, associations under this model are a mixture of two different types (Thornton *et al.* 2013). First, associations may be due to tagging a causal marker whose effect size is small, implying a sufficiently small effect on fitness, allowing the mutation to reach intermediate frequency ($>5\%$ in the population). The second class of association is due to noncausative mutations in linkage disequilibrium (LD) with causal markers. These "tagged" associations tend to be rare, and of relatively large effect (Thornton *et al.* 2013). Under this model, "missing heritability" arises from a combination of allelic heterogeneity, and a lack of power to identify risk variants.

Under the model of noncomplementing mutations, regions harboring risk alleles show a statistical signature of a large number of markers with single-marker p-values approaching, but still below, a genome-wide significance threshold (Thornton *et al.* 2013). These latter authors further showed that, under this model, the excess of significant markers (ESM) test, a permutation-based regional association test, had more power to detect a causal gene region in typical GWAS data than single marker methods, and many popular region-based tests (Thornton *et al.* 2013), even for GWAS containing only common markers (MAF >0.05). Although the test statistic of the ESM test is inspired by order statistics, under the permutation procedure for evaluating statistical significance, it is equivalent to the rank truncated product (RTP) of p-values (Dudbridge and Koeleman 2003). This equivalence was not initially recognized by Thornton *et al.* (2013). Multiple variations on the RTP exist to address issues related to correlation between p-values (De la Cruz *et al.* 2010), and the need to specify a truncation threshold (Yu *et al.* 2009). Although the RTP test has been used recently to obtain pathway- or gene-level associations in GWAS, and other, genomic applications (Meyer *et al.* 2012; Brenner *et al.* 2013; Ahsan *et al.* 2014; Li *et al.* 2014; Lee *et al.* 2014; Arem *et al.* 2015; Lai *et al.* 2015), it is not widely used. Here, we demonstrate the utility of mining existing datasets with an RTP approach, which we call the ESM test from here on, and provide an efficient implementation that can perform genome-wide scans without the need to restrict only to coding regions.

GWAS data do not have sufficient coverage of rare variants for direct analysis, but the ESM test is a powerful tool for extracting useful information despite this fact. Here we perform an empirical analysis of the performance of the ESM test on the Wellcome Trust Case Control Consortium (WTCCC) GWAS data set (Wellcome *et al.* 2007). We chose this dataset to determine the empirical efficacy of the ESM test because the dataset is well-characterized and easy to obtain. In addition, the choice of a dataset without substantial rare variant coverage, allows us to show that the ESM test has the power to detect the slight differences in allele frequencies between cases and control at common neutral markers, which is predicted by RALE models. We discover four novel gene regions that contribute to complex disease variation not detected in the original study, and propose that the ESM test is even better-suited to data sets that employ more modern, denser, SNP chips.

MATERIALS AND METHODS

Dataset

Data were obtained from the Wellcome Trust Case Control Consortium (http://www.wtccc.org.uk/), and are as described in Wellcome *et al.* (2007). Briefly, we obtained ~ 2000 cases for each of seven diseases, and a set of ~ 3000 shared controls typed on an Affymetrix 500K SNP chip. Diseases included in the dataset are Bipolar Disorder (BD), Coronary Artery Disease (CAD), Hypertension (HT), Chron's disease (IBD), Rheumatoid Arthritis (RA), Type 1 Diabetes(T1D), and Type 2 Diabetes (T2D). Case and control samples are obtained from across Great Britain. Control samples contain two subgroups: ~ 1500 individuals come from the 1958 British Birth Cohort (1958BC), and ~ 1500 belong to the national UK Blood Services donor pool (NBS).

Data preprocessing

The raw WTCCC data were formatted for use in PLINK 1.90a (Purcell *et al.* 2007). Single nucleotide polymorphisms (SNPs) listed in the WTCCC genotype file by their Affymetrix identification were translated into RefSNP (rsID) with the Affymetrix chip annotations. The SNP identifications and chromosome positions were updated to the most recent dbSNP Build 144. The SNP and individual exclusions lists provided were applied, and only genotyping calls with quality score over 0.9 were included.

Basic association and permutation

The basic single marker association test was executed with the PLINK 1.90a command *–assoc*. A total of N permuted single marker p-values are obtained from PLINK!1.90a by specifying –mperm = N. We take $N = 2\times 10^6$ permutations, such that the resolution of our permutation p-value is $\frac{1}{N} = 0.5\times 10^{-6}$, which can allow us to establish a region as genome-wide significant below a marginal p-value threshold of $\alpha \leq 1e-6$. We stored the observed association p-values, the permuted association p-values, and the R^2 between each marker (from plink *–ld* command) into HDF5 file format for use in the ESM test.

Excess of significant markers test

We implement the ESM test as described in Thornton *et al.* (2013). The test is a permutation based variation of rank truncated Fisher's combined p-value method using a null hypothesis based on order statistics. The test statistic is the sum of the differences between the observed and expected $-\log_{10}(p)$. However the expected value under the null is the same for each permutation and thus the statistic is equivalent to the sum of observed $-\log_{10}(p)$, *i.e.*, the RTP. For a set of m markers, the expected p-value, under the null model of no association, of the i^{th} most significant marker is $\frac{i}{m}$. Let $\mathbf{Y}$ be a vector of length, m, containing the observed $-\log_{10}(p)$, sorted in order of decreasing significance, from the single marker association test. Then the ESM test statistic is defined to be:

$$\text{ESM} = \sum_{i=1}^{m}\left(Y_i + \log_{10}\left(\frac{i}{m}\right)\right)$$

For each region, we calculate the ESM test statistic based for the observed data, and for each permutation of the data. For a given region, let the set of ESM test statistics be $\text{ESM}_j : j = 0, \ldots, N$, such that ESM_0

Table 1 New Associations: regions with ESM test $p \leq$ 1e-6 with no corresponding hit from Wellcome *et al.* (2007) are reported below

Disease	Chr	Position (Mb)	Gene Region	Source
CAD	7	80.78–80.88	SEMA3C	This analysis
CAD	7	129.993–130.123	ZC3HC1/KLHDC10	(Erbilgin *et al.* 2013)
T1D, RA	22	37.096–37.203	IL2RB	(Plagnol *et al.* 2011; Eleftherohorinou *et al.* 2011; Okada *et al.* 2014; Chimusa *et al.* 2014)
IBD	1	172.872–172.983	FASLG/TNFSF18	(Franke *et al.* 2010; Jostins *et al.* 2012; Dubois *et al.* 2010)

Three out of four regions contain corresponding hits in the NHGRI GWAS database not due to Wellcome *et al.* (2007) or were otherwise previously indicated in the particular disease as cited in the source column above. One region is novel based on our analysis, and overlaps with a biologically plausible gene SEMA3C. CAD, coronary artery disease; T1D, type 1 diabetes; RA, rheumatoid arthritis; IBD, Chron's disease.

is the observed value and the rest are calculated from permuted data. Then, the p-value for that region is:

$$p = \frac{\sum_{j=1}^{N} I(j)}{N}$$

where,

$$I(j) = \begin{cases} 1 & \text{ESM}_j \geq \text{ESM}_0 \\ 0 & \text{ESM}_j < \text{ESM}_0 \end{cases}$$

We performed the ESM test using a two-stage sliding window approach. Using 100-kb windows, we performed a genome scan with a jump size of 50 kb, with $m = 25$. The effect of changing m was explored in Thornton *et al.* (2013), and the choice of 25 was based on average SNP density in the WTCCC data. Within each region, we filtered markers based on LD, taking only SNPs whose R^2 was less than 0.2; always removing the SNP with the greater chromosomal position. While choosing this particular LD pruning rule is arbitrary, it prevents the introduction of bias due to selecting SNPs based on association significance. Regions that contained a marginally significant hit, with ESM p-values less than 1e–04, were rescanned using a finer (1 kb) jump size. The code for implementing the test can obtained at from github: https://github.com/ThorntonLab/ESMtest. Contiguous genomic regions that contain windows reaching genome-wide significance at $\alpha \leq$ 1e–6 were taken and explored for functional annotations. This significance threshold results in a predicted genome-wide Type-1 error rate of approximately 0.06; the mean (across diseases) number of total windows analyzed is 58,724, and, thus, the idealized type-1 error rate is $58,724 \times 1e-6 = 0.0587$. However, this estimate is quite conservative because the windows are spatially auto-correlated across the genome, making the effective number of tests performed much lower than the number of windows analyzed.

Intersection with other GWAS data

Significant regions were initially queried against the NHGRI GWAS database (http://www.ebi.ac.uk/gwas/). Regions were classified as being potentially novel if there were no significant SNPs in the NHGRI GWAS database for the specific disease whose genomic position fell within the boundaries of the region. Regions containing significant SNPs in the NHGRI GWAS database that were not contributed by the Wellcome Trust were also taken for further analysis. The regions were queried against gene and transcript annotations in human reference genome GRCh38 using the R package biomaRt (Durinck *et al.* 2005, 2009). The resulting gene and transcript annotations were manually curated for novelty and functional relevance.

Data availability

Data were obtained from the Wellcome Trust Case Control Consortium (http://www.wtccc.org.uk/).

RESULTS

We implement the ESM test as a sliding-window genome-wide scan for significant regions; we use 100 kb windows and 2 million permutations to reach genome-wide significance at an empirical $p \leq$ 1e-6. Region-/set-based methods result in far fewer tests than single-marker methods. By analyzing 60,000 windows with a marginal $\alpha \leq$ 1e-6, our genome-wide type 1 error rate was roughly 0.06; this estimate is conservative because the windows are not independent, and thus we effectively performed fewer tests than is suggested by the number of windows analyzed. Permutation procedures on genomic datasets are notoriously computationally expensive, and are thus typically avoided, despite their appealing statistical properties. With this in mind, we developed an efficient and freely available computational pipeline to implement the ESM test, which relies on new software and PLINK 1.90a (Purcell *et al.* 2007) (see *Materials and Methods*). The pipeline leverages PLINK's fast permutation procedures for single marker association tests, stores the data in I/O optimized HDF5 file format, and performs the test. Our analysis recapitulates most, but not all, of the associations established in the standard analysis of Wellcome *et al.* (2007) and finds new associations demonstrating that the ESM test is an excellent candidate for application in addition to standard methods.

Overlap between the ESM test and standard analysis

The majority of the regions found in Wellcome *et al.* (2007) that showed strong associations with case-control status were also significant under the ESM test. In Wellcome *et al.* (2007), the standard 1–df χ^2 test resulted in 21 regions showing strong association signals ($p \leq$ 5e-7). Supplemental Material, Table S1 shows that 18 of these regions also have an ESM test $p \leq$ 1e–6. Of the three regions that do not reach genome-wide significance under the ESM test, two have p-values between 1e–4 and 1e–6 (Table S2). In particular, multiple windows containing rs2542151, the main SNP reported for region chr18:12.77–12.92(Mb) in association with inflammatory bowel disease, reach ESM p =9e–6. A third SNP, rs420259, in region chr16:23.38-23.7(Mb) reported in association with bipolar disorder by Wellcome *et al.* (2007) did not replicate in other studies (Tung *et al.* 2011), and the region does not show strong association via the ESM test. Applying the SKAT (Wu *et al.* 2011; Lee *et al.* 2012) test to the same genomic windows results in less overlap with the WTCCC results (Table S4 and Table S5). Some of the regions not deemed significant by SKAT have been validated in other studies and can be viewed as false negatives. The ESM test has fewer false negatives. Because SKAT is not a permutation-based test, it is orders of magnitude faster computationally. However, our concern should focus primarily on getting better answers within the constraints of what is tractable. The ESM test is computationally feasible (Figure S2), and is shown here to give useful results. When we look at the overlaps and differences between the results of the ESM

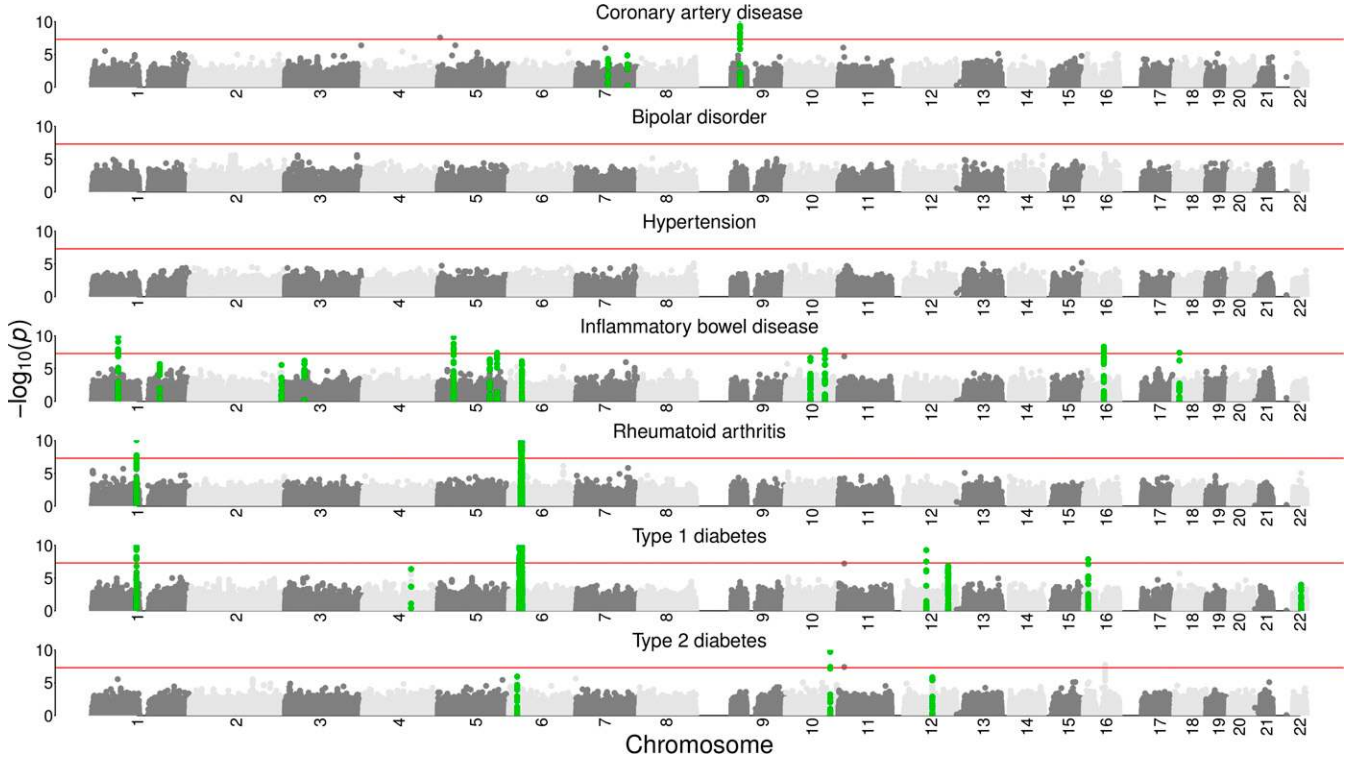

Figure 1 Manhattan plots with ESM significant regions highlighted. Single marker $-log_{10}(p)$*p*-values *vs.* chromosomal position (BP) for all seven diseases analyzed, with SNPs in ESM significant (ESM $p \leq$ 1e–6) regions highlighted in green. Horizontal lines are placed at $-log_{10}(p) = 8$ to illustrate the typical single marker genome-wide significance threshold. SNP clusters that are highlighted in green, but do not contain a single genome-wide significant SNP, are reported as novel.

test and the single marker test, we make two important observations. First, the ESM test has the power to detect genomic regions in association with disease status. Second, because there are regions that are only identified by either the ESM test, or the single marker test, we should view these methods as complementary. The second point is conceptually important, but computationally trivial because one has to do a single marker test to serve as the input to the ESM test. The suggested workflow is essentially as follows: run the single marker test, run the ESM test, analyze both results separately, and then observe their union and intersection.

Strong associations replicated in independent datasets

Table 1 shows that the ESM test identifies four genomic regions that were not significant in the original WTCCC single-marker based analysis. Three out of four of these regions have since been associated with disease statuses in independent studies published in the years following the introduction of the WTCCC analysis (Table 1). These subsequent independent studies all leveraged datasets employing larger case/control panels and/or more densely genotyped SNPs than were originally used in Wellcome *et al.* (2007). Published simulations suggest that the ESM test should accrue additional benefits when used on datasets with improved genotyping (see Figure 3 in Thornton *et al.* 2013). In contrast, applying SKAT (Wu *et al.* 2011; Lee *et al.* 2012) to these same data and genomic windows was less promising. Although SKAT finds three significant regions that are not significant with a single marker test (Table S3), only two have support in studies, and no completely novel candidate genes are found. The number of new results is not significantly different between the ESM test and SKAT, but there does appear to be a qualitative difference in the level of plausibility. However, at present we cannot rule out differences in optimal approach to partitioning the genome, or differences in the type of signal detected in explaining the observed differences in ESM and SKAT results. Overall, three of the four novel associations identified using the ESM test are replicated, providing empirical support that the ESM test can detect novel true positive associations, even in relatively small data sets. We briefly describe the known biological significance of these three genomic regions below.

The region chr7:129.99-130.12(Mb) is strongly associated with coronary artery disease (CAD) (Figure 1 and Table 1). This region overlaps two genes: *ZC3HC1* and *KLHDC10*. A missense mutation in ZC3HC1, which is also a *cis*-eQTL for *KLHDC10*, has been previously associated with CAD (Erbilgin *et al.* 2013). Neither gene currently has a clearly understood role in the etiology of CAD. The region chr22:37.09-37.21(Mb), containing *IL2RB*, is associated with type-1 diabetes (T1D) (Figure 1 and Table 1). This region was nominally associated with rheumatoid arthritis (RA) by the WTCCC, but not with T1D. *IL2RB* has been associated with both diseases in multiple studies (Plagnol *et al.* 2011; Eleftherohorinou *et al.* 2011; Okada *et al.* 2014; Chimusa *et al.* 2014). Epidemiological associations with immune related genes like *IL2RB* have motivated many important basic and clinical research studies (Pozzilli *et al.* 2015). Finally, we find an intergenic region, chr1:172.87-172.99(Mb), which contains SNPs previously associated with inflammatory bowel disease (IBD) (Franke *et al.* 2010; Jostins *et al.* 2012) and Celiac Disease (Dubois *et al.* 2010), to be associated with IBD (Figure 1 and Table 1). Both nearby genes, *TNFSF18* and *FASLG*, are part of the immunologically important TNF superfamily. The presence of putatively active regulatory elements within this associated region (Figure S1), supports the association between variation in

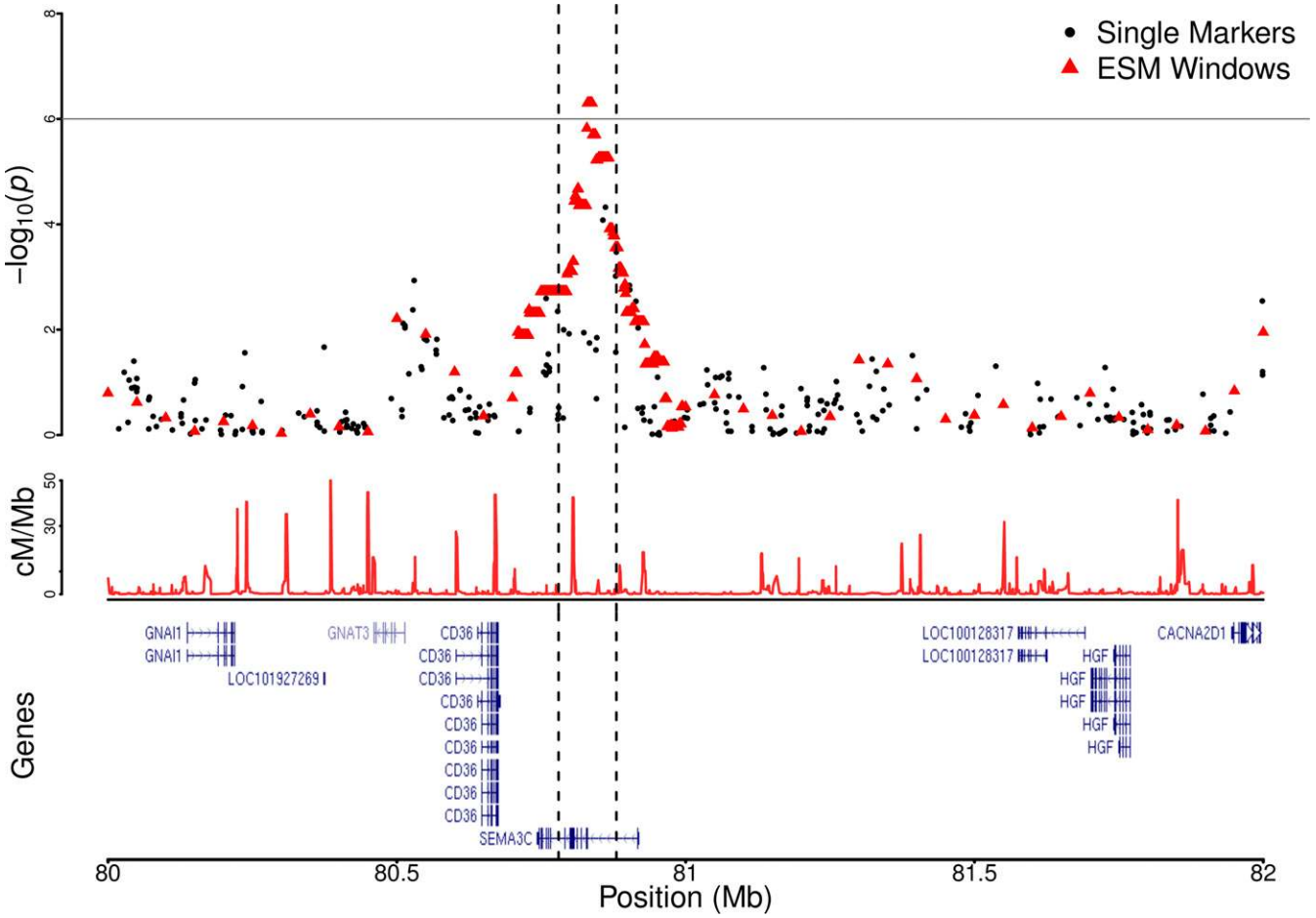

Figure 2 Region plot for SEMA3C hit. The top panel contains single marker (black points) and ESM test (red triangles) $-log_{10}(p)$-values for coronary artery disease *vs.* chromosomal position in the region chr7:80-82 (Mb). Each ESM test point is plotted at the midpoint of a genomic window to which that $-log_{10}(p)$-values corresponds. The single 100 kb ESM significant (ESM $p \leq$ 1e–6) region chr7:80.78-80.88 (Mb) is demarcated by vertical dashed lines, and the horizontal lines are placed at $-log_{10}(p) = 6$ to indicate the ESM test significance threshold. The middle panel contains the recombination rate in cM/Mb obtained from HapMap througout the same region. The lower panel shows the refseq gene UCSC genome browser track for the region.

regulatory sequences and common diseases (Maurano *et al.* 2012; Mathelier *et al.* 2015).

Novel association: SEMA3C

The ESM test finds one additional novel region, not shown to be of genome-wide significance in any study to date, showing strong association with CAD: chr7:8.08-8.09(Mb) (Table 1 and Figure 1). The only known protein-coding gene in this region is *SEMA3C* (Figure 2). A single SNP (rs4236644) in *SEMA3C* reached marginal significance (p =2e–6) in a meta-analysis of GWAS for total serum bilirubin levels (Johnson *et al.* 2009). *SEMA3C* is a secreted neurovascular guiding molecule that has a number of developmental functions, and plays a role in cardiovascular development during embyrogenesis (Püschel *et al.* 1995; Feiner *et al.* 2001). Certain congenital heart diseases are attributed to disregulation of *SEMA3C*, and its associated receptor *PLXNA2* (Kodo *et al.* 2009). *SEMA3C* is also an adipokine indicated in extracellular changes during white adipose tissue hypertrophy in human obesity (Mejhert *et al.* 2013). In total, *SEMA3C* is a plausible candidate gene driving the observed ESM signal. However, we should note that the nearby (0.5 Mb away) gene *CD36* is associated with heart-disease-related traits, including response to blood lipid drugs (Frazier-Wood *et al.* 2012), platelet count, and HDL cholesterol in African Americans (Qayyum *et al.* 2012; Coram *et al.* 2013). Although Figure 2 demonstrates lower support for *CD36*, its presence could be driving the association with *SEMA3C* through long-range LD. Alternatively, the presence of *CD36* might reflect the typical spatial clustering of functionally related genes found in many organisms (Hurst *et al.* 2004). Overall, the association of *SEMA3C* with CAD is consistent with its known physiological function in the development of the heart, and thus makes it an intriguing candidate for future studies.

DISCUSSION

The power of the ESM test is highlighted by the fact that it can identify novel, biologically plausible associations in an approximately 10-yr-old data set that has been highly studied. We provide open-source software implementing the test, which can be applied to GWAS data in PLINK .ped/.bed file format. As a caveat, although the test is simple, performing millions of permutations on GWAS data sets is computationally intensive. Individual-level genotype data are a requirement of the ESM test. The test cannot be applied to summary statistics from case/control studies. If it is applied to data with greater SNP coverage across the

genome, a finer-scale sliding window may be desirable, requiring more permutations to keep Type-1 errors low. Nevertheless, simulations suggest that the power of the ESM test will increase significantly when the test is applied to data sets that have employed more modern higher density SNP chips (Thornton *et al.* 2013). False positives due to LD between markers is often a concern for region-based analysis, although it has been shown that using permutation does adequately address the impact of LD on variations of Fisher's combined *p*-value (Moskvina *et al.* 2012; Alves and Yu 2014). However, when SNP pruning is applied, as it is here, to reduce the maximum pairwise correlation to 0.2, the effect is predicted to be quite insignificant (Alves and Yu 2014). This agrees with the observation from Thornton *et al.* (2013) that the ESM test did not result in any false positives under neutral simulations.

We find that using rank truncated product methods in conjunction with single-marker analysis yields an approximately 20% gain in power over single-marker analysis alone, as illustrated by the finding of four new results on top of the preexisting 21 results from the standard method. It is clear to see the potential benefit of applying the ESM test in this way to all of the existing GWAS data. Given the extent of GWAS data currently in existence, it is conceivable that a broad application of the ESM test would establish thousands of new associations. An additional benefit of a broad application the ESM test is the opportunity to validate hits in new datasets with older ones, as we demonstrated here.

A key limitation of region/SNP-set based tests in general, including rank truncated product methods, is that one cannot simply validate a single or small set of markers in a second panel. It is instead necessary to do deep genotyping of a candidate region in an independent panel in order to gain a perspective on the genetic variation present in the associated region. A corollary is that the lack of simple single SNP markers makes the estimation of effect sizes and variance explained by a detected gene region difficult; this problem should be a focus of future studies. Using existing data, rank truncated product methods have power to detect new associations between genomic regions and disease. Notably, the development of more powerful region-based tests seems likely. The ESM test was designed to detect an association signal in case/control panels under a particular gene action model, and a small range of population genetic scenarios. Recent work (Moutsianas *et al.* 2015) demonstrates that predictions from simulation studies regarding performance of region-based tests are impacted by various model details. Thus, future research should focus on the behavior of association tests under various models of gene action and demography.

ACKNOWLEDGMENTS

We are thankful to Harry Mangalam, Joseph Farran, and Adam Brenner, for administering the University of California, Irvine High-Performance Computing cluster. We are thankful to J.J. Emerson, Kirk Lohmueller, Michael Zwick, and Andy Clark for helpful comments. We also thank Mahul Chakraborty and James Baldwin-Brown for software testing. This work was supported by NIH grant R01-GM115564 to KRT, and NIH grant R01-GM115562 to ADL. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE-1321846. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

LITERATURE CITED

Ahsan, H., J. Halpern, M. G. Kibriya, B. L. Pierce, and L. Tong, *et al.*, 2014 A genome-wide association study of early-onset breast cancer identifies PFKM as a novel breast cancer gene and supports a common genetic spectrum for breast cancer at any age. Cancer Epidemiol. Biomarkers Prev. 23: 658–669.

Alves, G., and Y.-K. Yu, 2014 Accuracy evaluation of the unified P-value from combining correlated P-values. PLoS One 9: e91225.

Arem, H., K. Yu, X. Xiong, K. Moy, N. D. Freedman *et al.*, 2015 Vitamin D metabolic pathway genes and pancreatic cancer risk. PLoS One 10: e0117574.

Auer, P. L., M. Nalls, J. F. Meschia, B. B. Worrall, W. T. Longstreth *et al.*, 2015 Rare and coding region genetic variants associated with risk of ischemic stroke: the NHLBI Exome sequence project. JAMA Neurol. 72: 781–788.

Brenner, A. V., G. Neta, E. M. Sturgis, R. M. Pfeiffer, A. Hutchinson *et al.*, 2013 Common single nucleotide polymorphisms in genes related to immune function and risk of papillary thyroid cancer. PLoS One 8: e57243.

Chimusa, E. R., N. Zaitlen, M. Daya, M. Möller, P. D. van Helden *et al.*, 2014 Genome-wide association study of ancestry-specific TB risk in the South African Coloured population. Hum. Mol. Genet. 23: 796–809.

Cirulli, E. T., and D. B. Goldstein, 2010 Uncovering the roles of rare variants in common disease through whole-genome sequencing. Nat. Rev. Genet. 11: 415–425.

Coram, M. A., Q. Duan, T. J. Hoffmann, T. Thornton, J. W. Knowles *et al.*, 2013 Genome-wide characterization of shared and distinct genetic components that influence blood lipid levels in ethnically diverse human populations. Am. J. Hum. Genet. 92: 904–916.

Cruchaga, C., C. M. Karch, S. C. Jin, B. A. Benitez, Y. Cai *et al.*, 2014 Rare coding variants in the phospholipase D3 gene confer risk for Alzheimer's disease. Nature 505: 550–554.

De la Cruz, O., X. Wen, B. Ke, M. Song, and D. L. Nicolae, 2010 Gene, region and pathway level analyses in whole-genome studies. Genet. Epidemiol. 34: 222–231.

Dudbridge, F., and B. P. C. Koeleman, 2003 Rank truncated product of P-values, with application to genomewide association scans. Genet. Epidemiol. 25: 360–366.

Dupuis, J., C. Langenberg, I. Prokopenko, R. Saxena, N. Soranzo *et al.*, 2010 New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. Nat. Genet. 42: 105–116.

Durinck, S., Y. Moreau, A. Kasprzyk, S. Davis, B. De Moor *et al.*, 2005 BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. Bioinformatics 21: 3439–3440.

Durinck, S., P. T. Spellman, E. Birney, and W. Huber, 2009 Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Nat. Protoc. 4: 1184–1191.

Eichler, E. E., J. Flint, G. Gibson, A. Kong, S. M. Leal *et al.*, 2010 Missing heritability and strategies for finding the underlying causes of complex disease. Nat. Rev. Genet. 11: 446–450.

Eleftherohorinou, H., C. J. Hoggart, V. J. Wright, M. Levin, and L. J. M. Coin, 2011 Pathway-driven gene stability selection of two rheumatoid arthritis GWAS identifies and validates new susceptibility genes in receptor mediated signalling pathways. Hum. Mol. Genet. 20: 3494–3506.

Erbilgin, A., M. Civelek, C. E. Romanoski, C. Pan, R. Hagopian *et al.*, 2013 Identification of CAD candidate genes in GWAS loci and their expression in vascular cells. J. Lipid Res. 54: 1894–1905.

Feiner, L., A. L. Webber, C. B. Brown, M. M. Lu, L. Jia *et al.*, 2001 Targeted disruption of semaphorin 3C leads to persistent truncus arteriosus and aortic arch interruption. Development 128: 3061–3070.

Fisher, R. A., 1930 *The Genetical Theory Of Natural Selection*. Clarendon Press, Oxford, UK.

Franke, A., D. P. B. McGovern, J. C. Barrett, K. Wang, G. L. Radford-Smith *et al.*, 2010 Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. Nat. Genet. 42: 1118–1125.

Frazier-Wood, A. C., S. Aslibekyan, I. B. Borecki, P. N. Hopkins, C.-Q. Lai *et al.*, 2012 Genome-wide association study indicates variants associated with insulin signaling and inflammation mediate lipoprotein responses to fenofibrate. Pharmacogenet. Genomics 22: 750–757.

Gibson, G., 2012 Rare and common variants: twenty arguments. Nat. Rev. Genet. 13: 135–145.
Haldane, J. B. S., 1927 A mathematical theory of natural and artificial selection, Part V: selection and mutation. Math. Proc. Camb. Philos. Soc. 23: 838–844.
Hurst, L. D., C. Pál, and M. J. Lercher, 2004 The evolutionary dynamics of eukaryotic gene order. Nat. Rev. Genet. 5: 299–310.
Huyghe, J. R., A. U. Jackson, M. P. Fogarty, M. L. Buchkovich, A. Stančáková *et al.*, 2013 Exome array analysis identifies new loci and low-frequency variants influencing insulin processing and secretion. Nat. Genet. 45: 197–201.
Johansen, C. T., J. Wang, M. B. Lanktree, H. Cao, D. Adam *et al.*, 2011 Mutation skew in genes identified by genome-wide association study of hypertriglyceridemia. Nat. Genet. 42: 684–687.
Johnson, A. D., M. Kavousi, A. V. Smith, M. H. Chen, A. Dehghan *et al.*, 2009 Genome-wide association meta-analysis for total serum bilirubin levels. Hum. Mol. Genet. 18: 2700–2710.
Johnston, H. R., Y. Hu, and D. J. Cutler, 2015 Population genetics identifies challenges in analyzing rare variants. Genet. Epidemiol. 39: 145–148.
Jostins, L., S. Ripke, R. K. Weersma, R. H. Duerr, D. P. McGovern *et al.*, 2012 Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. Nature 491: 119–124.
Kodo, K., T. Nishizawa, M. Furutani, S. Arai, E. Yamamura *et al.*, 2009 GATA6 mutations cause human cardiac outflow tract defects by disrupting semaphorin-plexin signaling. Proc. Natl. Acad. Sci. USA 106: 13933–13938.
Lai, Y.-C., C.-F. Kao, M.-L. Lu, H.-C. Chen, P.-Y. Chen *et al.*, 2015 Investigation of associations between NR1D1, RORA and RORB genes and bipolar disorder. PLoS One 10: e0121245.
Lee, E., J. Luo, Y.-C. Su, J. P. Lewinger, F. R. Schumacher *et al.*, 2014 Hormone metabolism pathway genes and mammographic density change after quitting estrogen and progestin combined hormone therapy in the California Teachers Study. Breast Cancer Res. 16: 477.
Lee, S., M. J. Emond, M. J. Bamshad, K. C. Barnes, M. J. Rieder *et al.*, 2012 Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. Am. J. Hum. Genet. 91: 224–237.
Li, W.-Q., R. M. Pfeiffer, P. L. Hyland, J. Shi, F. Gu *et al.*, 2014 Genetic polymorphisms in the 9p21 region associated with risk of multiple cancers. Carcinogenesis 35: 2698–2705.
Manolio, T. A., F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff *et al.*, 2009 Finding the missing heritability of complex diseases. Nature 461: 747–753.
Mathelier, A., W. Shi, and W. W. Wasserman, 2015 Identification of altered cis-regulatory elements in human disease. Trends Genet. 31: 67–76.
Maurano, M. T., R. Humbert, E. Rynes, R. E. Thurman, E. Haugen *et al.*, 2012 Systematic localization of common disease-associated variation in regulatory DNA. Science 337: 1190–1195.
McClellan, J., and M.-C. King, 2010 Genetic heterogeneity in human disease. Cell 141: 210–217.
Mejhert, N., F. Wilfling, D. Esteve, J. Galitzky, V. Pellegrinelli *et al.*, 2013 Semaphorin 3C is a novel adipokine linked to extracellular matrix composition. Diabetologia 56: 1792–1801.
Meyer, T. E., L. W. Chu, Q. Li, K. Yu, P. S. Rosenberg *et al.*, 2012 The association between inflammation-related genes and serum androgen levels in men: the prostate, lung, colorectal, and ovarian study. Prostate 72: 65–71.
Moskvina, V., K. M. Schmidt, A. Vedernikov, M. J. Owen, N. Craddock *et al.*, 2012 Permutation-based approaches do not adequately allow for linkage disequilibrium in gene-wide multi-locus association analysis. Eur. J. Hum. Genet. 20: 890–896.
Moutsianas, L., V. Agarwala, C. Fuchsberger, J. Flannick, M. A. Rivas *et al.*, 2015 The power of gene-based rare variant methods to detect disease-associated variation and test hypotheses about complex disease. PLoS Genet. 11: e1005165.
Naser, S. A., M. Arce, A. Khaja, M. Fernandez, N. Naser *et al.*, 2012 Role of ATG16L, NOD2 and IL23R in Crohn's disease pathogenesis. World J. Gastroenterol. 18: 412–424.
Nelson, M. R., D. Wegmann, M. G. Ehm, D. Kessner, P. St Jean *et al.*, 2012 An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. Science 337: 100–104.
Okada, Y., D. Wu, G. Trynka, T. Raj, C. Terao *et al.*, 2014 Genetics of rheumatoid arthritis contributes to biology and drug discovery. Nature 506: 376–381.
Plagnol, V., J. M. M. Howson, D. J. Smyth, N. Walker, J. P. Hafler *et al.*, 2011 Genome-wide association analysis of autoantibody positivity in type 1 diabetes cases. PLoS Genet. 7: e1002216.
Pozzilli, P., E. Maddaloni, and R. Buzzetti, 2015 Combination immunotherapies for type 1 diabetes mellitus. Nat. Rev. Endocrinol. 11: 289–297.
Prescott, N. J., K. M. Dominy, M. Kubo, C. M. Lewis, S. A. Fisher *et al.*, 2010 Independent and population-specific association of risk variants at the IRGM locus with Crohn's disease. Hum. Mol. Genet. 19: 1828–1839.
Prescott, N. J., B. Lehne, K. Stone, J. C. Lee, K. Taylor *et al.*, 2015 Pooled sequencing of 531 genes in inflammatory bowel disease identifies an associated rare variant in BTNL2 and implicates other immune related genes. PLoS Genet. 11: e1004955.
Pritchard, J. K., 2001 Are rare variants responsible for susceptibility to complex diseases? Am. J. Hum. Genet. 69: 124–137.
Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. A. R. Ferreira *et al.*, 2007 PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 81: 559–575.
Purcell, S. M., J. L. Moran, M. Fromer, D. Ruderfer, N. Solovieff *et al.*, 2014 A polygenic burden of rare disruptive mutations in schizophrenia. Nature 506: 185–190.
Püschel, A. W., R. H. Adams, and H. Betz, 1995 Murine semaphorin D/collapsin is a member of a diverse gene family and creates domains inhibitory for axonal extension. Neuron 14: 941–948.
Qayyum, R., B. M. Snively, E. Ziv, M. A. Nalls, Y. Liu *et al.*, 2012 A meta-analysis and genome-wide association study of platelet count and mean platelet volume in african americans. PLoS Genet. 8: e1002491.
Robinson, M. R., N. R. Wray, and P. M. Visscher, 2014 Explaining additional genetic variation in complex traits. Trends Genet. 30: 124–132.
Sham, P. C., and S. M. Purcell, 2014 Statistical power and significance testing in large-scale genetic studies. Nat. Rev. Genet. 15: 335–346.
Spencer, C. C. A., Z. Su, P. Donnelly, and J. Marchini, 2009 Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip. PLoS Genet. 5: e1000477.
Thornton, K. R., A. J. Foran, and A. D. Long, 2013 Properties and modeling of GWAS when complex disease risk is due to non-complementing, deleterious mutations in genes of large effect. PLoS Genet. 9: e1003258.
Todd, J. A., N. M. Walker, J. D. Cooper, D. J. Smyth, K. Downes *et al.*, 2007 Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes. Nat. Genet. 39: 857–864.
Tung, J. Y., C. B. Do, D. A. Hinds, A. K. Kiefer, J. M. Macpherson *et al.*, 2011 Efficient replication of over 180 genetic associations with self-reported medical data. PLoS One 6: e23473.
Visscher, P. M., W. G. Hill, and N. R. Wray, 2008 Heritability in the genomics era–concepts and misconceptions. Nat. Rev. Genet. 9: 255–266.
Visscher, P. M., M. A. Brown, M. I. McCarthy, and J. Yang, 2012a Five years of GWAS discovery. Am. J. Hum. Genet. 90: 7–24.
Visscher, P. M., M. E. Goddard, E. M. Derks, and N. R. Wray, 2012b Evidence-based psychiatric genetics, AKA the false dichotomy between common and rare variant hypotheses. Mol. Psychiatry 17: 474–485.
Weersma, R. K., P. C. F. Stokkers, I. Cleynen, S. C. S. Wolfkamp, L. Henckaerts *et al.*, 2009 Confirmation of multiple Crohn's disease susceptibility loci in a large Dutch-Belgian cohort. Am. J. Gastroenterol. 104: 630–638.

Wei, W.-H., G. Hemani, and C. S. Haley, 2014 Detecting epistasis in human complex traits. Nat. Rev. Genet. 15: 722–733.

Wellcome, T., T. Case, and C. Consortium, 2007 Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447: 661–678.

Welter, D., J. MacArthur, J. Morales, T. Burdett, P. Hall *et al.*, 2014 The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res. 42: 1001–1006.

Wessel, J., and M. O. Goodarzi, 2015 Low-frequency and rare exome chip variants associate with fasting glucose and type 2 diabetes susceptibility. Nat. Commun. 6: 5897.

Wu, M. C., S. Lee, T. Cai, Y. Li, M. Boehnke *et al.*, 2011 Rare-variant association testing for sequencing data with the sequence kernel association test. Am. J. Hum. Genet. 89: 82–93.

Yu, K., Q. Li, A. W. Bergen, R. M. Pfeiffer, P. S. Rosenberg *et al.*, 2009 Pathway analysis by adaptive combination of P-values. Genet. Epidemiol. 33: 700–709.

Zuk, O., E. Hechter, S. R. Sunyaev, and E. S. Lander, 2012 The mystery of missing heritability: genetic interactions create phantom heritability. Proc. Natl. Acad. Sci. USA 109: 1193–1198.

Communicating editor: R. Cantor